BJA

British Journal of Anaesthesia, xxx (xxx): xxx (xxxx)

doi: 10.1016/j.bja.2021.09.025
Advance Access Publication Date: xxx
Review Article

REVIEW ARTICLE

Artificial intelligence for mechanical ventilation: systematic review of design, reporting standards, and bias

Jack Gallifant[1,*], Joe Zhang[2,3,†], Maria del Pilar Arias Lopez[4,†], Tingting Zhu[5], Luigi Camporota[1,2], Leo A. Celi[6,7,8,*] and Federico Formenti[1,9,10]

[1]Centre for Human and Applied Physiological Sciences, School of Basic and Medical Biosciences, King's College London, London, UK, [2]Department of Adult Critical Care, Guy's and St Thomas' NHS Foundation Trust, King's Health Partners, London, UK, [3]Institute of Global Health Innovation, Imperial College London, London, UK, [4]SATI-Q Program, Argentine Society of Intensive Care, Buenos Aires, Argentina, [5]Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford, UK, [6]Laboratory for Computational Physiology, Massachusetts Institute of Technology, Cambridge, MA, USA, [7]Division of Pulmonary, Critical Care and Sleep Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA, [8]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA, USA, [9]Nuffield Division of Anaesthetics, University of Oxford, Oxford, UK and [10]Department of Biomechanics, University of Nebraska Omaha, Omaha, NE, USA

*Corresponding authors. E-mails: jack.gallifant@kcl.ac.uk, lceli@mit.edu

†These authors contributed equally.

Abstract

Background: Artificial intelligence (AI) has the potential to personalise mechanical ventilation strategies for patients with respiratory failure. However, current methodological deficiencies could limit clinical impact. We identified common limitations and propose potential solutions to facilitate translation of AI to mechanical ventilation of patients.
Methods: A systematic review was conducted in MEDLINE, Embase, and PubMed Central to February 2021. Studies investigating the application of AI to patients undergoing mechanical ventilation were included. Algorithm design and adherence to reporting standards were assessed with a rubric combining published guidelines, satisfying the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis [TRIPOD] statement. Risk of bias was assessed by using the Prediction model Risk Of Bias ASsessment Tool (PROBAST), and correspondence with authors to assess data and code availability.
Results: Our search identified 1,342 studies, of which 95 were included: 84 had single-centre, retrospective study design, with only one randomised controlled trial. Access to data sets and code was severely limited (unavailable in 85% and 87% of studies, respectively). On request, data and code were made available from 12 and 10 authors, respectively, from a list of 54 studies published in the last 5 yr. Ethnicity was frequently under-reported 18/95 (19%), as was model calibration 17/95 (18%). The risk of bias was high in 89% (85/95) of the studies, especially because of analysis bias.
Conclusions: Development of algorithms should involve prospective and external validation, with greater code and data availability to improve confidence in and translation of this promising approach.
Trial registration number: PROSPERO − CRD42021225918.

Keywords: artificial intelligence; bias; critical care; decision support; mechanical ventilation respiratory failure

Received: 30 July 2021; **Accepted:** 27 September 2021

Editor's key points

- This systematic review was conducted to identify common limitations and potential solutions to the translation of artificial intelligence to mechanical ventilation of patients.
- Of 1,342 studies identified, 95 studies were included, most of which were single-centre retrospective studies with limited access to data sets and code, and high risk of bias.
- Artificial intelligence applied to mechanical ventilation has limited external validation and model calibration, with substantial risk of bias, significant gaps in reporting, and poor code and data availability.
- Future development of algorithms should involve prospective and external validation, with rigorous adherence to standards, and code and data availability to facilitate translation of data science into improved approaches to mechanical ventilation.

Invasive mechanical ventilation is a vital supportive therapy for critically ill patients with respiratory failure.[1] These patients are heterogeneous in terms of disease aetiology, lung pathology, and respiratory mechanics;[2–4] ventilation strategies based on current guidelines do not guarantee lung protection to individual patients, and inappropriate settings can increase morbidity and mortality.[5–7] The best available evidence is based on results from a few clinical trials, where outcomes are identified at a population level and significant heterogeneity exists between individual patients.[8,9] The risk of applying population data to single patients has been highlighted by secondary analyses of the trial data.[10] However, the experimental evidence to guide personalised ventilation strategies is lacking.

Artificial intelligence (AI) could offer a solution in the pursuit towards personalised mechanical ventilation, capitalising on the widespread use of electronic monitoring and recording in high-income countries and the associated wealth of data generated during mechanical ventilation of intraoperative and ICU patients. The capacity of AI to predict sepsis,[11] circulatory failure,[12] and patient mortality[13] demonstrates the potential for further applications in the anaesthesia and critical care settings. However, the current AI landscape is increasingly acknowledged as incomplete, with frequent concerns regarding AI reproducibility and generalisability,[14–16] and methodological limitations during algorithm design and validation can limit transferability and clinical translation. Reporting guidelines have been developed to improve standards of publication and to reduce the potential for bias. However, these parameters have not been evaluated for AI specifically in mechanical ventilation.[17]

We aimed to identify common methodological limitations and propose changes in publication standards that would enable translation of AI algorithms into clinical practice to support mechanical ventilation. Therefore, we conducted a systematic review of the AI literature as applied to mechanical ventilation, evaluating adherence to the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement, risk of bias using the Prediction model Risk Of Bias ASsessment Tool (PROBAST), algorithm design using a novel rubric, and the availability of both data and code.[18,19]

Methods

The systematic review protocol was registered with the online International Prospective Register of Systematic Reviews database (CRD42021225918) before search execution; deviations are reported in Supplementary file 1. This study was created in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines,[20] with a checklist displayed in Supplementary file 2. No institutional ethical approval was required.

Study identification and inclusion criteria

A comprehensive search strategy was performed using free text and MeSH terms of various forms of the keywords 'artificial intelligence' and 'mechanical ventilation'; the full search strategy is outlined in Supplementary file 3. Three major electronic databases (MEDLINE, Embase, and PubMed) were searched from database creation until February 26, 2021, with no additional filters used. Additional articles were identified from references of included studies. Inclusion and exclusion criteria are detailed in Table 1.

Study selection and data extraction

All texts were screened by one author (JG), and subsets were also screened by a second author (FF, MDPAL, or JZ) so that each text was screened twice independently. Disagreements emerged from this independent screening were resolved by consensus. Data were systematically extracted from each study into a predesigned spreadsheet and analysed *post hoc* using pivot tables.

Adherence to standards

Reporting quality was assessed using a rubric generated by MIT Critical Data, led by the Laboratory for Computational Physiology, combining published guidelines (Supplementary file 4). This approach satisfies the TRIPOD statement,[19] 'a checklist that aims to increase transparency of publications developing predictive models for medicine'.[21] Further emphasis is provided on the impact of generalisability and assessment of the availability of data and code, which are fundamental for reproducibility in this discipline.

In addition, the authors of included studies published since January 1, 2016 (n=54) for which code or data were not publicly available were e-mailed for access. The time frame was limited to the past 5 yr based upon the significant increase in publications over this period and to identify current barriers to sharing data sets, not to determine whether data are kept for prolonged periods of time. Risk of bias was assessed using PROBAST.[18] The overall group adherence was calculated as a percentage, and adherence groups created based upon reporting of 0–33.3%, 33.4–66.6%, or 66.7–100%. Descriptive statistics were calculated as mean (standard deviation [SD]) for continuous variables, and frequencies and rates for categorical data. No formal quantitative synthesis was performed because of the heterogeneity in outcomes studied, models used, and reporting levels.

AI maturity

There is no agreed upon framework to determine the maturity of clinical AI literature.[22,23] We assessed the AI maturity of each study using the US Food and Drug Administration (FDA)

Table 1 Inclusion and exclusion criteria for study selection. *Invasive mechanical ventilation is defined as a mechanism of respiratory support delivered to a patient with a tracheal tube. Paediatric patients were included to provide a comprehensive assessment of the literature. †Artificial intelligence algorithms were defined as computational programmes with the capacity to learn, evaluate their own performance, and update their rules, to facilitate a prediction output. ‡Human clinicians had to be equivalent level to board certification/completion of specialty training to be considered expert. AI, artificial intelligence; NA, not applicable.

	Inclusion criteria	Exclusion criteria
Participants	Participants (adult and paediatric) undergoing invasive mechanical ventilation*	Non-human participants (animals or modelling data generated algorithmically)
Intervention	Assess application of AI† to patient data Investigation of outcomes directly related to mechanical ventilation	Application of AI to inform internal ventilator operation, increase clinical imaging techniques resolution, or identify clinical anatomy Primary use of neural signals or genetic/molecular markers as candidate predictors Models predicting successful intubation or investigating obstructive sleep apnoea or noninvasive ventilation
Comparator (if presented)	Clinicians‡ or previously validated models	NA
Study type	Published peer-reviewed scientific reports in English until the search date	Review articles, commentaries, letters, editorials, and other informal publication types
Outcome	Model performance Patient-related outcomes (e.g. mortality and ventilation requirement)	NA
Context	Hospitals with critical care facilities that manage patients undergoing invasive mechanical ventilation Data generated in this setting, in the context of routine clinical care, or generated in the context of prospective research	NA

guidelines and expert opinion.[24] Studies were categorised into the following groups: (i) 'maths into algorithm' (authors propose new algorithmic construct with some feasibility testing), (ii) 'algorithm into model' (authors develop a new model based on retrospective data with some prospective feasibility testing), (iii) 'model into device' (validation of a previously developed model, ideally against an existing gold standard), and (iv) 'device into practice' (prospective clinical evaluation of deployed devices).

Results

Study selection

We identified 1,342 studies, including 434 duplicates, resulting in 908 studies for which abstracts were screened (Fig. 1). After abstract and full-text screening, a total of 95 studies were included in the systematic review; a full list of included and excluded studies can be found in Supplementary file 5.

Study characteristics

Several studies were single-centre studies (53/94; 56%) (one not reported), primarily with a retrospective design (84/95; 88%), few were conducted prospectively (11/95; 12%), and only one RCT. The five countries with the highest number of published studies were the USA (34/95; 36%), Spain (13/95; 14%), China (7/95; 7%), Brazil (5/95; 5%), and UK (5/95; 5%), together representing 67% of the included studies; one study did not define the country of the studied population.

Three studies recruited neonatal patients (mean age 22 weeks); the remaining studies had a population age (mean [SD]) of 61 [8] yr (64/95 reported). A total of 58 studies (61%) reported the sex of participants (61% male; 39% female). Only 18/95 (19%) studies reported at least part of the ethnic distribution of the cohort: overall, the mean distribution [SD] was 57 [17]% White, 26 [20]% Black, 29 [23]% Asian, and 5 [2]% Hispanic; studies may have reported 81% for a single ethnicity and no details of other ethnicities, so percentages do not sum to 100%.

The number of publications by year increased significantly from 2016 to 2020, with 11 studies publishing new models in the first 2 months of 2021 alone (Fig. 2). However, progression towards maturity was limited, as no 'device into practice' technology was identified, despite the increasing rate of 'algorithm into model' studies published in recent years.

Adherence to standards

Table 2 presents percentage adherence of relevant outcomes; notably, there were very limited freely available code and data, available in only 13% and 15%, respectively. We e-mailed 54 authors for access to code and data: one e-mail was automatically returned because of a non-existent account; 12 and 10 were able to provide access respectively to code and data on request; three were not able to comply because of ethical or regulatory reasons; and 37 did not respond.

Risk of bias

Overall, 10 studies were classified as low risk, whereas the remaining 85 were classified as high risk of bias (Fig. 3). The

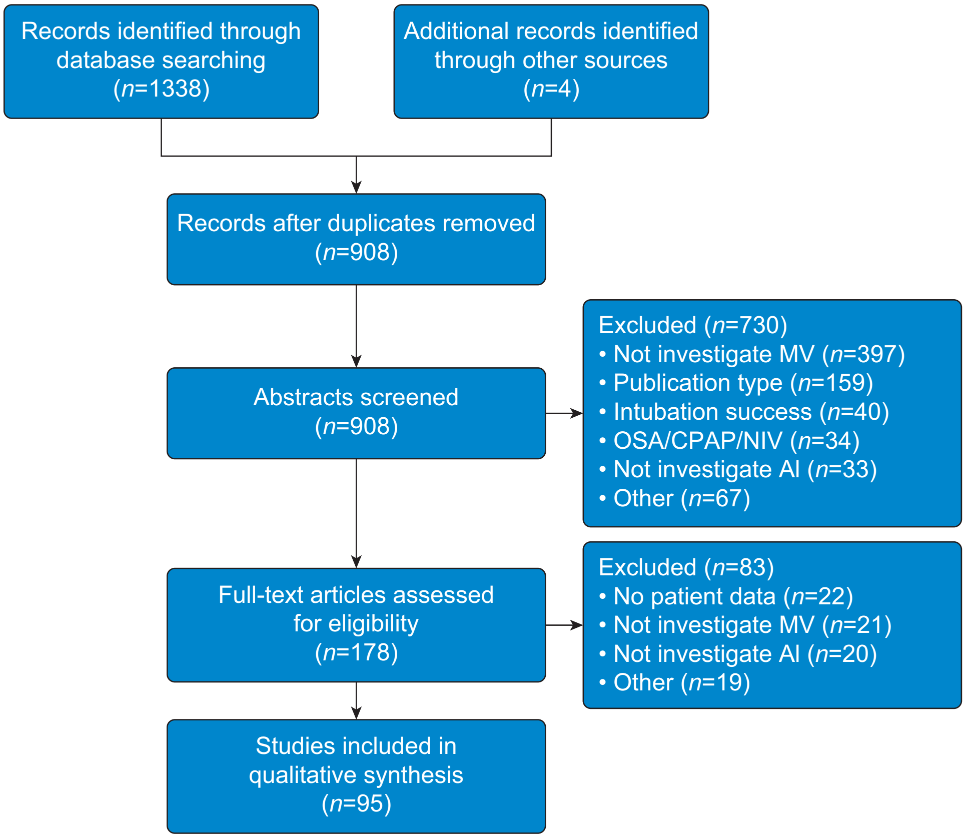
Records identified through database searching (*n*=1338)
Additional records identified through other sources (*n*=4)
Records after duplicates removed (*n*=908)
Abstracts screened (*n*=908)
Excluded (*n*=730)
• Not investigate MV (*n*=397)
• Publication type (*n*=159)
• Intubation success (*n*=40)
• OSA/CPAP/NIV (*n*=34)
• Not investigate AI (*n*=33)
• Other (*n*=67)
Full-text articles assessed for eligibility (*n*=178)
Excluded (*n*=83)
• No patient data (*n*=22)
• Not investigate MV (*n*=21)
• Not investigate AI (*n*=20)
• Other (*n*=19)
Studies included in qualitative synthesis (*n*=95)

Fig 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow chart of study records. The full list of screened papers is available in Supplementary file 5. AI, artificial intelligence; MV, mechanical ventilation; NIV, noninvasive ventilation; OSA, obstructive sleep apnoea.

majority were deemed low risk in participants (84/95), predictors (88/95), and outcome analysis (95/95), but 85/95 were deemed high risk of bias in the analysis section. This high risk of bias was attributable to inadequate reporting of missing data handling (40/95) and insufficient detail on model performance (79/95), where the majority failed to mention calibration (78/95).

Clinical outcomes

The four most published prediction outcome categories were predicting weaning success (23/95; 24%), predicting commencement of mechanical ventilation (i. e. within 24 h) (22/95; 23%), predicting a complication in ventilated patients (18/95; 19%), and detecting patient–ventilator asynchrony (12/95; 13%); only 6% (6/95) reported AI to inform a clinical decision support system. The most common models used were neural networks (30/95; 32%), decision trees (28/95; 29%), and clustering algorithms (10/99; 11%).

The majority of studies (75/95; 79%) did not compare the results of the model developed or validated with either clinician performance, a previously validated model, or standard index (e.g. early warning score). The same majority conducted analysis on cohort results, without any individualised approach like bedside support systems.

Discussion

To our knowledge, this is the first systematic review conducted specifically in the field of AI for mechanical ventilation. The exponentially growing body of AI literature has largely attempted to predict onset or aid mechanical ventilation weaning, with less focus on clinical decision support. Methodological deficiencies were frequent and changes in practice are necessary to ensure development of a robust evidence base in this emerging field.

External validation is critical to protect patients and trust in AI algorithms

A significant majority of retrospective analysis was observed, with only one RCT.[25] This finding has been observed in other domains and limits the potential for translation to clinical care.[26,27]

Despite greatly improved internal validation techniques, there is still limited external validation, and few multicentre studies exist, which limit both confidence in the validity, and applicability to different geographical regions. A disproportionately large contribution comes from a few countries (e. g. USA; 34/95) and in some cases cities (e. g. Boston, MA; 8/95). This skewed geographical distribution limits the algorithms' applicability to demographically and genotypically different

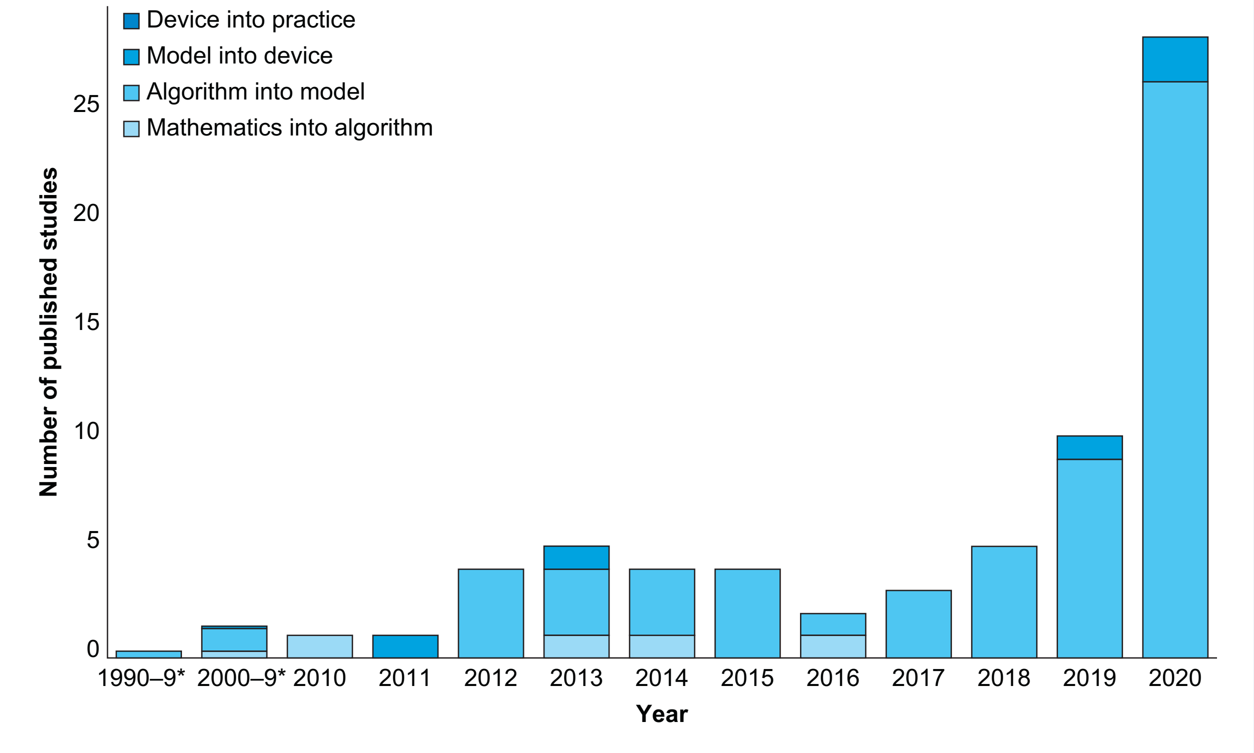

Fig 2. Temporal distribution illustrating the number and maturity of studies included in this review. The average number of publications per year is presented for two decades (1990–9* and 2000–9*) because of the limited number of relevant publications in that period; the highest number of publications in any 1 yr during that period was three (2002 and 2009). There has been a significant rise in the number of artificial intelligence algorithms being used to develop models applied to mechanical ventilation in the last 5 yr; 55 since January 2016. However, there has been no consistent shift towards device creation and subsequent deployment and evaluation in clinical practice.

populations, and to practices that differ between health systems, such as criteria for ICU admission and intubation, which may not be accounted for in the algorithms.[16] These factors will affect the distribution of features in the training and validation cohorts, introducing collider bias and leading to spurious associations that find their way into the algorithms.[28] A ‘collider’ refers to an independent third variable caused by an exposure and outcome; collider bias is the distorted association between the exposure and outcome when controlling a collider.[29] This has important implications for the sampling methods used in data science and in database formation. A potential benefit of benchmark data sets that represent diverse patient cohorts is that they could allow for evaluation of generalisability and recalibration of algorithms.[30]

Table 2 Adherence to reporting standards. The respective Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) reference is displayed in round brackets and the percentage adherence in square brackets. NA, not applicable.

Bottom tertile (0–33.3)	Middle tertile (33.4–66.6)	Upper tertile (66.7–100)
External validation [21%]	Outcome comparison (14a) [65%]	Medical context (3a) [100%]
Outliers (NA) [20%]	Exclusion criteria (5b) [64%]	Objectives (3b) [100%]
Ethnicity split (13b) [19%]	Abstract (2) [63%]	Study design (4a) [100%]
Calibration (16) [18%]	Study dates (4b) [62%]	Prediction outcome (6a) [100%]
Data freely available (21) [15%]	Clinically relevant (16) [59%]	Inclusion criteria (5b) [99%]
Code freely available (21) [13%]	Missing data (9) [58%]	Setting (5a) [98%]
Protocol (21) [4%]	Sex split (13b) [56%]	Sample size (8) [98%]
		Data flow/summary (13a) [98%]
		Defined predictors (7a) [97%]
		Discrimination (16) [97%]
		Internal validation (10b) [95%]
		Data pre-processing (10a) [93%]
		Funding (22) [84%]
		Full prediction model (15a) [75%]
		Mean age of sample (13b) [67%]

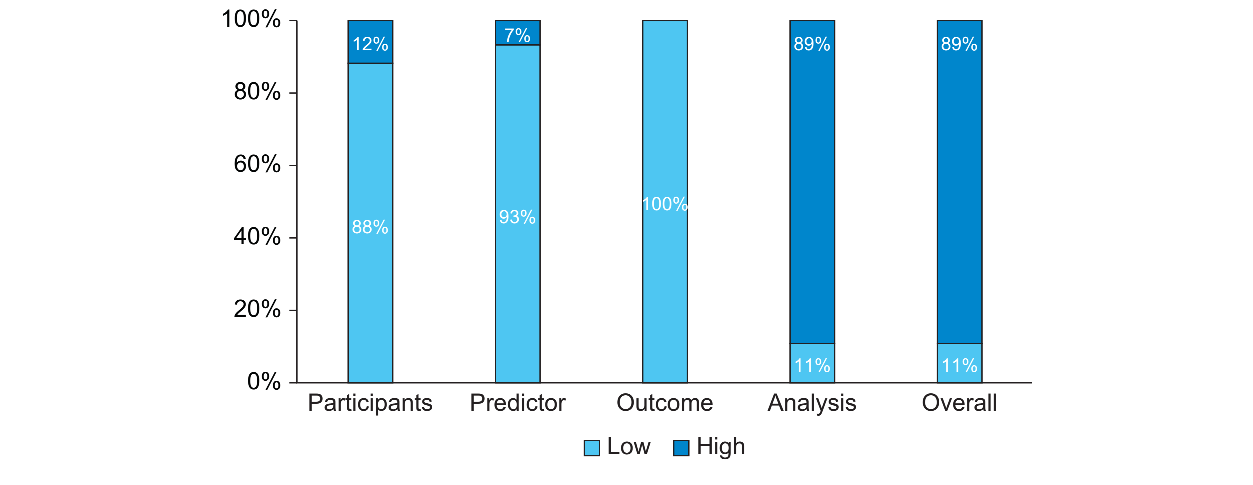

Fig 3. Risk of bias assessed using Prediction model Risk Of Bias ASsessment Tool (PROBAST) reporting standards for all included studies. Low indicates the percentage of studies that did not receive any high-risk rating for a particular category. High denotes those that achieved a high-risk rating in at least one criterion for a particular category.

Ethnicity was largely under-reported, making comparison of algorithm performance when applied to different ethnicities impossible, and potentially contributing to disparities once implemented.[31–33] If sufficient prospective and/or external validation has demonstrated benefits to patients and/or clinicians, generalisability may be falsely assumed, but it is vital to ensure algorithm performance is continuously monitored, and especially its effect on health disparities.[16]

Calibration translates model discrimination into patient-level outcomes

Model discrimination was universally reported with the most common index being area under the receiver operating characteristic curve. It is important that clinicians work with model developers to present clinically meaningful and relevant metrics, such as precision and recall. This is highlighted by differences in ideal sensitivity and specificity level for weaning patients undergoing mechanical ventilation and for predicting mortality. However, model calibration was consistently under-reported. Calibration has important implications for patient-level (rather than population-level) performance, and is necessary for monitoring drifts after deployment, even within the same centre. A model may have near-perfect discrimination, but if it is poorly calibrated, clinicians would be reasonably reluctant to adopt it.

The most investigated outcomes were predicting weaning success and ventilator complications, a disproportion possibly attributable to the clear-cut nature of the outcomes. There was a predominance of neural network and random forest algorithms, where the former may represent relatively superior discrimination power capabilities at the expense of interpretability. All except three studies provided an explanation of algorithm performance and attempted to increase interpretability, but the use of tools, such as feature importance or SHapley Additive exPlanations, was under-utilised (3/95).[34]

Translation of interpretable models to the bedside remains a significant challenge, where collaboration is required to refocus data science researchers towards clinically meaningful and translatable problems. The gap between AI model development and subsequent device implementation was evident (Fig. 2) with a lack of device creation or subsequent assessment in clinical practice, despite significant increase in model development. These findings are typical of the broader AI literature, with a wide discrepancy between the large number of AI publications in 2019 (n=12,422) and the number of new useable algorithms receiving approval from the FDA for clinical use in 2020 (n=130).[35,36]

Moving beyond the ‘black box’ of algorithm development

Our systematic review identified a dearth of publicly available data and code. Upon direct request, few authors (12/54) were able to share data and/or code, with the rest of responders unable to share data and/or code because of institutional ethics, risk of patient data being released, and/or confidentiality over the code to be submitted for separate peer-reviewed publication or patent applications, even if the majority of articles stated ‘Data will be made available on request’, or similar. This outcome highlights the need for a system leading to internationally achievable anonymisation and curation of data, with their dissemination to the research community. Code availability should become the rule rather the exception, and anonymised data should at least be available on request, or preferably stored in publicly available and secure sites.

The importance of access to well-curated data is fundamental for algorithm development, as is the access process transparency. A notable feature in the evaluated studies is the absence of reporting of how missing data and outliers were managed. Missing data are inevitable in electronic health records, and the disproportionate loss of data between different subgroups or patients can skew the modelling, similarly to attrition bias in randomised studies.

The COVID pandemic highlighted the limited ability of centres to rapidly aggregate clinical data, resulting in a number of ‘small, incompatible data’ in contrast with the acclaimed ‘big, standardised data’. There is a clear need for standardised structuring, formatting, and curating of collected data to facilitate generalisability and multicentre evaluation,[37] requiring close collaboration between clinicians, data scientists, and AI engineers throughout the life cycle of an algorithm. Reciprocal familiarisation of clinicians with AI, and of data scientists and AI engineers with clinical medicine, is required for an efficient progress in the field.

Overall, there is a significant risk of bias in the reporting methodology. This bias does not appear to be specific to mechanical ventilation, as it appears to be a consistent problem throughout the AI in healthcare domain.[38]

Future implications

To develop more robust prediction models, cross-disciplinary multicentre collaboration and improved reporting of model performance are required. These improvements will require data and code availability to become standard in the field to prioritise reproducibility over novelty. The goal of personalising ventilation strategy, and more generally healthcare, to individual patients will require integration of bidirectional data flow: from patients to algorithms for training, and from trained algorithms to patients, further highlighting the importance of merging data science with clinical medicine. The primary challenge in this flow of data is the availability of large multicentre, interoperable critical care data sets that contain the rich, longitudinal data required for producing generalisable models.

In the short term, making anonymised patient data publicly available for international collaborative research could contribute to a larger data set derived from diverse patient cohorts, redrawing the unequal medical knowledge map, with benefits reaching beyond national borders and into regions with limited resources. A key benefit of a system that could regularly provide new data is the potential for regular verification and re-evaluation of patient care, an iterative process that is critical for AI algorithm implementation. This process would require a new approach to governance, where current institutional policies may not consider potential for multiregional collaboration, or solutions for rapid turnover of data.

Long term, the proliferation of interoperable electronic health systems and data standards will result in the ability to accumulate higher-quality data at source and wider adoption of federated approaches for machine learning.[39] It could also support improved ICU care by providing evidence from more comprehensive data sets to be applied to the context of the local clinical scenario and to complement local expertise, reducing the bias inevitably associated with anecdotal experience. This approach could also enable the opportunity to assess the diversity of management styles for evidence-based protocols. True interoperability may be some time off, where coordination between stakeholders and improved data access will be key steps.[40]

Study limitations

The guidelines chosen to assess the studies (TRIPOD and PROBAST) were designed for prediction models only, so their application to assess other AI algorithms may pose limitations, for example where a few studies may have been overlooked. Our elaboration to these guidelines and development of AI maturity framework were based upon expert opinion and have not been tested before. The PROBAST risk of bias does contain a subjective component, and borderline results may affect overall interpretation. The decision on which studies directly investigated mechanical ventilation and on what classified as AI was associated with subjective interpretation and definitions.

Conclusions

Currently, AI applied to mechanical ventilation has limited external validation and model calibration, with substantial risk of bias, significant gaps in reporting, and poor code and data availability. Rigorous adherence to standards, and broad changes in our approach to data and reproducibility, will facilitate translation of algorithmic data science into deployable tools and improved patient care.[41]

Authors’ contributions

Conceptualisation: JG, FF, LAC.
Data analysis: JG, MPAL, LC, FF.
Interpretation: JG, JZ, MPAL, TZ, LAC, FF.
Preparation of paper: JG, FF.
Critical revision and approval of paper: all authors.
All authors approved the final version of the paper and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. All persons designated as authors qualify for authorship, and all those who qualify for authorship are listed. All authors had final responsibility for the decision to submit for publication.

Declarations of interest

FF reports grants from the National Institute for Health Research (UK), the National Association of Academic Anaesthetists, and the Wellcome EPSRC Centre for Medical Engineering; he is one of the volunteer directors of OxVent, a joint-venture social enterprise between Oxford University and King’s College London for mechanical ventilation. All other authors declare no competing conflicts.

Funding

US National Institutes of Health (NIBIB R01 EB017205) to LAC; National Institute for Health Research Invention for Innovation (#200681) to FF; National Institute of Academic Anaesthesia (NIAA19R108) to FF; Wellcome Trust to JZ; Engineering for Development Research Fellowship provided by the Royal Academy of Engineering to TZ.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bja.2021.09.025.

References

1. Fan E, Brodie D, Slutsky AS. Acute respiratory distress syndrome: advances in diagnosis and treatment. *JAMA* 2018; **319**: 698–710
2. Wilson JG, Calfee CS. ARDS subphenotypes: understanding a heterogeneous syndrome. *Crit Care* 2020; **24**: 102

3. Luo L, Shaver CM, Zhao Z, et al. Clinical predictors of hospital mortality differ between direct and indirect ARDS. *Chest* 2017; **151**: 755–63
4. Chiumello D, Busana M, Coppola S, et al. Physiological and quantitative CT-scan characterization of COVID-19 and typical ARDS: a matched cohort study. *Intensive Care Med* 2020; **46**: 2187–96
5. Slutsky AS, Ranieri VM. Ventilator-induced lung injury. *N Engl J Med* 2013; **369**: 2126–36
6. Cavalcanti AB, Suzumura ÉA, Laranjeira LN, et al. Effect of lung recruitment and titrated positive end-expiratory pressure (PEEP) vs low PEEP on mortality in patients with acute respiratory distress syndrome: a randomized clinical trial. *JAMA* 2017; **318**: 1335–45
7. Constantin JM, Jabaudon M, Lefrant JY, et al. Personalised mechanical ventilation tailored to lung morphology versus low positive end-expiratory pressure for patients with acute respiratory distress syndrome in France (the LIVE study): a multicentre, single-blind, randomised controlled trial. *Lancet Respir Med* 2019; **7**: 870–80
8. Brower RG, Matthay MA, Morris A, et al., Acute Respiratory Distress Syndrome Network. Ventilation with lower tidal volumes as compared with traditional tidal volumes for acute lung injury and the acute respiratory distress syndrome. *N Engl J Med* 2000; **342**: 1301–8
9. The National Heart, Lung, and Blood Institute ARDS Clinical Trials Network. Higher versus lower positive end-expiratory pressures in patients with the acute respiratory distress syndrome. *N Engl J Med* 2004; **351**: 327–36
10. Goligher EC, Costa ELV, Yarnell CJ, et al. Effect of lowering V_T on mortality in acute respiratory distress syndrome varies with respiratory system elastance. *Am J Respir Crit Care Med* 2021; **203**: 1378–85
11. Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nat Med* 2018; **24**: 1716–20
12. Hyland SL, Faltys M, Hüser M, et al. Early prediction of circulatory failure in the intensive care unit using machine learning. *Nat Med* 2020; **26**: 364–73
13. Pirracchio R, Petersen ML, Carone M, Rigon MR, Chevret S, van der Laan MJ. Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): a population-based study. *Lancet Respir Med* 2015; **3**: 42–52
14. Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. *Sci Data* 2016; **3**: 160035
15. Celi LA, Mark RG, Stone DJ, Montgomery RA. "Big data" in the intensive care unit. Closing the data loop. *Am J Respir Crit Care Med* 2013; **187**: 1157–60
16. Futoma J, Simons M, Panch T, Doshi-Velez F, Celi LA. The myth of generalisability in clinical research and machine learning in health care. *Lancet Digit Health* 2020; **2**: e489–92
17. Vasey B, Clifton DA, Collins GS, et al. DECIDE-AI: new reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence. *Nat Med* 2021; **27**: 186–7
18. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med* 2019; **170**: 51–8
19. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual Prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ* 2015; **350**: g7594
20. Moher D, Liberati A, Tetzlaff J, Altman DG, The PG. Preferred reporting Items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med* 2009; **6**, e1000097
21. Moons KGM, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015; **162**: W1–73
22. van de Sande D, van Genderen ME, Huiskens J, Gommers D, van Bommel J. Moving from bytes to bedside: a systematic review on the use of artificial intelligence in the intensive care unit. *Intensive Care Med* 2021; **47**: 750–60
23. Komorowski M. Artificial intelligence in intensive care: are we there yet? *Intensive Care Med* 2019; **45**: 1298–300
24. US Food & Drug Administration. *Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD)*. discussion paper and request for feedback 2019. Available from: https://www.fda.gov/files/medical%20devices/published/US-FDA-Artificial-Intelligence-and-Machine-Learning-Discussion-Paper.pdf. [Accessed 10 July 2021]
25. Hsu JC, Chen YF, Chung WS, Tan TH, Chen T, Chiang JY. Clinical verification of a clinical decision support system for ventilator weaning. *Biomed Eng Online* 2013; **12**: S4
26. Nagendran M, Chen Y, Lovejoy CA, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ* 2020; **368**: m689
27. Roberts M, Driggs D, Thorpe M, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nat Mach Intell* 2021; **3**: 199–217
28. Charpignon M-L, Celi LA, Samuel MC. Who does the model learn from? *Lancet Digit Health* 2021; **3**: e275–6
29. Lee H, Aronson JK, Nunan D. Catalogue of bias collaboration. In: *Catalogue of bias*; 2019. Available from: https://catalogofbias.org/biases/collider-bias/. [Accessed 10 July 2021]
30. Stupple A, Singerman D, Celi LA. The reproducibility crisis in the age of digital medicine. *NPJ Digit Med* 2019; **2**: 2
31. Soto GJ, Martin GS, Gong MN. Healthcare disparities in critical illness. *Crit Care Med* 2013; **41**: 2784–93
32. White DB, Lo B. Mitigating inequities and saving lives with ICU triage during the COVID-19 pandemic. *Am J Respir Crit Care Med* 2021; **203**: 287–95
33. Editorial. Race representation matters in cancer care. *Lancet Digit Health* 2021; **3**: e408
34. Lundberg SML, Su-In L. A unified approach to interpreting model predictions. *Advances In Neural Information Processing Systems*. 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4768-4777.

35. Benjamens S, Dhunnoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *NPJ Digit Med* 2020; **3**: 118
36. Wu E, Wu K, Daneshjou R, Ouyang D, Ho DE, Zou J. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. *Nat Med* 2021; **27**: 582–4
37. Lehne M, Sass J, Essenwanger A, Schepers J, Thun S. Why digital medicine depends on interoperability. *NPJ Digit Med* 2019; **2**: 79
38. Cirillo D, Catuara-Solarz S, Morey C, et al. Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. *NPJ Digit Med* 2020; **3**: 81
39. Rieke N, Hancox J, Li W, et al. The future of digital health with federated learning. *NPJ Digit Med* 2020; **3**: 119
40. Warren LR, Clarke J, Arora S, Darzi A. Improving data sharing between acute hospitals in England: an overview of health record system distribution and retrospective observational analysis of inter-hospital transitions of care. *BMJ Open* 2019; **9**, e031637
41. European Commission. *Laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts*. Directorate-General for Communications Networks, Content and Technology; 2021. Available from: https://www.parlament.gv.at/PAKT/EU/XXVII/EU/05/86/EU_58690/imfname_11061214.pdf. [Accessed 10 July 2021]

Handling editor: Hugh C Hemmings Jr